LEVEL II

TECHNICAL REPORT PTR-1060-78-6

A LITTLE LEARNING...: CONFIDENCE IN MULTICUE JUDGMENT TASKS

DECISION RESEARCH • A BRANCH OF PERCEPTRONICS

Baruch Fischhoff
Paul Slovic

June 1978

ADVANCED DECISION TECHNOLOGY PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

The objective of the Advanced Decision Technology Program is to develop and transfer to users in the Department of Defense advanced management technologies for decision making. These technologies are based upon research in the areas of decision analysis, the behavioral sciences and interactive computer graphics. The program is sponsored by the Cybernetics Technology Office of the Defense Advanced Research Projects Agency and technical progress is monitored by the Office of Naval Research – Engineering Psychology Programs. Participants in the program are:

Decisions and Designs, Incorporated
Harvard University
Perceptronics, Incorporated
Stanford Research Institute
Stanford University
The University of Southern California

Inquiries and comments with regard to this report should be addressed to:

Dr. Martin A. Tolcott
Director, Engineering Psychology Programs
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

or

Dr. Stephen J. Andriole
Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

(12)

TECHNICAL REPORT PTR-1060-78-6

A LITTLE LEARNING...: CONFIDENCE IN MULTICUE JUDGMENT TASKS

by

Baruch Fischhoff and Paul Slovic

Sponsored by

Defense Advanced Research Projects Agency
ARPA Order No. 3469

Under Subcontract from
Decisions and Designs, Incorporated

June 1978

DECISION RESEARCH
A BRANCH OF PERCEPTRONICS
1201 Oak Street
Eugene, Oregon 97401
(503) 485-2400

SUMMARY

Overview

A series of eight experiments investigated people's confidence in their ability to make a variety of judgments. Participants were almost uniformly overconfident in their abilities, even when warned of the difficulty of the tasks. Such overconfidence can have a very adverse effect on how information is recruited and analyzed in the making of decisions.

Background and Approach

A large component of any decision maker's job is to summarize complex ensembles of information into dichotomous judgments. On the basis of intelligence reports, it might be necessary to decide whether a particular set of maneuvers are exercises or the early stages of an attack. On the basis of personal impressions and reports, one might have to decide whether a particular officer is or is not competent. On the basis of prior experience, one might have to decide whether a recruit is ready for combat. An important aspect of such judgments is the degree of confidence that accompanies them. That confidence may determine whether more information will be gathered or whether an action will be taken. It may also determine whether that action will be tentative or restrained.

Earlier research in this program has found that overconfidence typifies most judgments studied. Those judgments were, however, generally restricted to confidence in general knowledge on a variety of unrelated tasks. In the present experiments, the participants assessed their confidence in a series of dichotomous judgments regarding one topic. Furthermore, they were given time to familiarize themselves with that topic and in some instances, given information relevant to their general level of ability on the task.

Findings and Implications

Without exception, the varied tasks used here were judged to be easier than was actually the case. Such overconfidence typified 80% of the participants in each study. Allowing participants to study a set of solved problems of the same type neither increased nor decreased overconfidence. A modest (but far from complete) reduction in overconfidence was effected by telling people that one task was virtually impossible.

From the results, it appeared that even minimal familiarity with a judgment task produces a great number of hypotheses regarding how it may be accomplished. These hypotheses are not tested properly; they are assumed to be correct--producing overconfidence.

Such overconfidence may lead to premature cessation of information gathering and to ineffective decision making. No generally effective way to combat it is available.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

Our thanks to Michael Enbar, Sol Fulero, Gerry Hanson and Mark Layman for help in conducting these experiments, to Nancy Collins and Peggy Roecker for typing and clerical assistance, and to Barb Combs for keeping our spirits up with encouraging words like "You guys ought to call that study 'hoax springs eternal.'" We also thank Sarah Lichtenstein for her comments on an earlier draft and William Graziano for bringing the introductory quotation to our attention.

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by Office of Naval Research under Contract N00014-78-C-0100 (ARPA Order No. 3469) under Subcontract 78-072-0722 from Decisions and Designs, Inc. to Perceptronics, Inc. Correspondence may be addressed to either author at Decision Research, A Branch of Perceptronics, 1201 Oak Street, Eugene, Oregon 97401.

1. INTRODUCTION

> The human understanding is of its own nature prone to suppose the existence of more order and regularity in the world than it finds. And though there be many things in nature which are singular and unmatched, yet it devises for them parallels and conjugates and relatives which do not exist.
>
> Bacon

Many tasks we face in life may be described as multi-cue discriminations. Using information from a number of variables, we make judgments such as adequate-inadequate, malignant-benign, fast-slow, or Democrat-Republican. What determines our confidence in our ability to make a particular kind of discrimination? One important cue is likely to be how well we seem to have been able to make such discriminations in the past. How well do we ascertain that ability? We should have the most realistic appraisal when we have gone through a concentrated series of trials in each of which we first make the required discrimination and then receive accurate outcome feedback, perhaps with instruction in why we did well or poorly (Hammond & Summer, 1972).

Such ideal conditions are, in most people's lives, quite rare. Typically, trials are so spread out that it is difficult to extract general discriminatory principles; feedback is ambiguous or so long in coming that we cannot remember exactly what our judgment was or how we made it; no one is around to instruct us, and so on. The opportunities for extracting the wrong amount of confidence, either too much or too little, are enormous.

One seemingly minor deviation from these ideal conditions is having concentrated trials with tasks and feedback presented simultaneously. For example, we might be presented a set of clinical profiles labeled "neurotic" or "psychotic" or race horses labeled "won" or "lost" or stocks labeled "rose" or "fell." We are to study these sets in order to determine how differently labeled cases differ and to assess our ability to make that discrimination when faced in the future with unlabeled cases.

The present experiments examine the appropriateness of assessments of discriminatory ability derived under such conditions. All subjects received sets of learning trials in which experience was concentrated and stimuli were presented in a clear, common format. For some subjects, the study stimuli were labeled (e.g., malignant or benign), for others, they were unlabeled. At first glance, it might seem as though subjects receiving labeled stimuli would be in the best position to appriase their discriminatory ability. We predicted, however, that provision of labels would mislead and produce unwarranted confidence in one's judgments.

At least three lines of evidence pointed in this direction. For one, Fischhoff (1975, 1977) has found that when people are told the outcomes to historical events, they overestimate the likelihood that they would have been able to predict those outcomes had they not been told; when told the answers to general knowledge questions, they overestimate how much they knew without being told. Apparently, once the outcome to an event or the answer to a question is reported, everything else known about that event or question is quickly reinterpreted to make a coherent whole of all relevant knowledge. People do not appreciate the extent of this reinterpretation and, as a result, exaggerate the extent to which they would

have been able to predict the answers, had they been asked. In a discrimination task, such a "knew-it-all-along" effect would lead people who have seen labeled trials to believe that they would have made more correct discriminations than would have been the case. It might also lead them to overestimate their ability to make such discriminations in the future.

The second line of evidence is anecdotal and may be found in methodological discussions of "correlational overkill" (Kunce, Cook & Miller, 1975) or the "degrees of freedom" problem (Campbell, 1975). Given a set of labeled cases and a sufficiently large set of characterizing attributes, one can always devise a rule predicting the labels from the attributes to any desired level of proficiency. In regression terms, by expanding the set of independent variables one can always find a set of predictors (or even one predictor) with any desired correlation with the independent variable. The price one pays for overfitting is, of course, shrinkage, failure of the predictive (or discriminatory) rule to "work" on a new sample of cases. The frequency and vehemence of the methodological warnings suggest that correlational overkill is a bias that is quite resistant to even extended professional training (Armstrong, 1975; Crask & Perreault, 1977; Hammer, 1974; Lewis-Beck, 1977).[1] The knew-it-all-along effect may be considered a form of overfitting by which attributes are selected, interpreted and highlighted so as to make the assigned labels seem obvious. Overconfidence in future discrimination tasks would arise if judges did not realize the extent to which they may have capitalized on chance when explaining the labels in the study sample.

Thirdly, the opportunity to study labeled examples may also lead to overconfidence in one's ability to make future discriminations by creating an illusion of control. As Langer (1975, 1977) has argued, people overestimate their future suc-

cess at tasks perceived to be dependent on skill (rather than luck). Furthermore, they tend to see an element of skill in situations that are governed by chance. Studying labeled examples might be expected to evoke undue feelings of skill (and control). These feelings would be augmented by hindsight effects and overfitting.

In order to see whether provision of labels with study examples induces overconfidence, we used a relatively small number of study examples (10-12), each of which was characterized by many attributes. Subjects' task was always to make a dichotomous discrimination on a subsequent set of unlabeled examples and indicate the probability of their choice being correct. The tasks were designed to appear difficult, but to be impossible.

In Experiment 1, for example, the task involved categorizing short handwriting samples as being written by either a European or an American. We predicted that allowing people to study a number of correctly labeled samples would increase their confidence in being able to make future discriminations without actually improving their ability. Control groups studying the same samples without labels should be equally proficient, but less confident.

2. EXPERIMENT 1 - HANDWRITING ANALYSIS

Method

Design. In Part I, every subject studied 10 handwriting specimens, five written by Americans and five by Europeans for a period of five minutes. For the labels group, these specimens were labeled correctly according to continent of origin. For the no-labels group, the specimens were unlabeled. In Part II, all subjects were given 10 additional specimens. For each, they were asked to make a best guess at the country of origin and to assess the probability that their guess was correct, using a probability from .50 to 1.00.

Stimuli. The 20 specimens used (10 European and 10 American) were selected from a set of 100 (50 European) collected by Dr. Lewis Goldberg in Eugene, Oregon and in The Netherlands. The criterion for inclusion was being correctly identified by between 40% and 60% of a sample of 20 student subjects in Eugene (mean percent correct = 52.3%). We believed that discrimination was impossible for these specimens and unlikely to improve with the minimal opportunity for learning offered the labels groups. The 20 specimens were randomly sorted into two sets of 10 (5 European; 5 American) in four different ways. Roughly one quarter of the subjects in each group received each sorting; half of these received each of the two sets in Part I, half in Part II. Thus, the 20 specimens used were presented in 8 different ways, in order to minimize the likelihood of using one particular combination with unusually good or poor transfer from Part I to Part II.

Instructions. Part I instructions read:

In this experiment, we are trying to determine whether people can distinguish between European and

American handwriting. You will see 10 cards. Each card will contain a simple handwritten sentence:

Mensa mea bona est

You are to judge whether each sentence was written by an American or a European.

Before you take this test, you will have an opportunity to study samples of handwriting. You will be given a page with ten [labeled--training group] samples, 5 American, and 5 European. You will have 5 minutes to study them prior to the test.

Part II instructions read:

Now that you have had a chance to examine the handwriting samples, you will have the opportunity to make some predictions. On the following pages, you will see some handwriting specimens. For each specimen, first indicate whether you think it was written by an American or European.

Second, decide what the probability is that your answer is correct. This probability can be any number from .5 to 1.0. It can be interpreted as your degree of certainty about the correctness of your answer. For example, if you respond that the probability is .6, it means that you believe that there are about 6 chances out of 10 that your answer is correct. A response of 1.0 means that you are absolutely certain that your answer is correct. A response of .5 means that your best guess is as likely to be right as wrong. Don't estimate any probability below .5, because you should always be picking the alternative that you think is more likely to be correct. Write your probability on the space provided on the answer sheet.

To repeat, this probability is a measure of your degree of certainty that your chosen alternative is the correct alternative. It is a number from .5 to 1.0, where .5 means complete uncertainty and 1.0 means complete certainty.

Subjects. A total of 52 paid subjects were recruited through an advertisement in the University of Oregon student paper. They were assigned to the two groups according to their preference for date and time at which the groups were scheduled. Subjects in subsequent experiments were recruited and assigned in the same way.

Results and Discussion

Table 1 presents the mean percent correct and mean probability judgment for subjects in each group. Subjects who saw the labels in Part I were more confident than subjects who did not (mean probability of .745 vs. .645). Unfortunately for the evaluation of our hypothesis, this increased confidence was highly justified. The minimal learning opportunity they received enabled labels subjects to identify correctly three quarters of the test specimens. Subjects without that little learning did little better than chance (53% correct) in Part II.

If subjects use the probability scale correctly (i.e., if they are "perfectly calibrated," see Lichtenstein & Fischhoff, 1977, or Lichtenstein, Fischhoff & Phillips, 1977), then their mean probability judgment should equal their percentage correct. By this criterion, the level of confidence of subjects in the labels condition was much more appropriate to their abilities than was that of the no-labels condition, which

Table 1

Performance (Percentage Correct) and Confidence (Mean Probability) in Part II

Experiment		Labels				No-Labels			
No.	Name	% Correct	Mean Probab.	Over/Under Confidence[a]	N	% Correct	Mean Probab.	Over/Under Confidence[a]	N
1	Handwriting	77.0	.745	-.025	22	53.3	.645	.112	30
2	Ulcers	76.3	.702	-.061	33	58.5	.599	.014	38
3	Stocks	49.3	.643	.150	38	44.0	.671	.229	25
4	Horseracing	41.5	.603	.188	46	39.1	.651	.260	42
5	Children's Drawings	54.1	.667	.126	47	52.3	.677	.154	45
6	Children's Drawings (discouraging instructions)	57.7	.631	.054	40	45.6	.627	.171	36

[a]Equals difference between mean probability and proportion correct. Negative sign indicates underconfidence.

showed considerable overconfidence. Thus, the labels group both knew more and had a better appreciation of how much they knew. This result fits a pattern reported by Lichtenstein and Fischhoff (1977), who found that the appropriateness of probability responses increases as percent correct increases from 50% to about 75% (above which it decreases).

3. EXPERIMENT 2 - ULCERS

Clearly, Experiment 1 did not provide an adequate test of the hypothesis that a worthless opportunity to learn an impossible task will lead people to be overconfident. The opportunity provided to labels subjects in Part I of Experiment 1 was much more useful than we imagined it would be.

Experiment 2 was designed to provide subjects with a completely unfamiliar task (and one that presumably could not be learned), diagnosing ulcers as malignant or benign on the basis of a small number of diagnostic signs. Cases were drawn from a study by Slovic, Rorer and Hoffman (1971) which discovered, among other things, substantial disagreement in diagnosis among the expert radiologists who served as subjects.

The seven diagnostic signs were the size of the ulcer (larger or smaller than 2 cm), its location (on or off the greater curvature) and the presence or absence of "extraluminality," "associating filling defect," "a regular contour," "a rugal pattern (i.e., radiating folds)" and "associated duodenal ulcer." No further explication of these signs was provided. Subjects saw eight examples in each of Part I and Part II. Those seen in Part I were either labeled benign or malignant or unlabeled. A total of 16 cases was used, divided into two sets, each of which appeared in Part I for half the subjects.

These were not actual cases, but artificial ones originally designed to be believable to a practicing radiologist. As a result, the actual diagnosis (the correct label) could only be guessed at. From screening large populations, several

of these seven signs have been found to have diagnostic validity. The 16 cases used here were found by Slovic et al. (1971) to have these valid signs pointing overwhelmingly toward one diagnosis (that used as the label).

Results

Much to our surprise (and chagrin), the pattern of Experiment 1 was repeated. As shown in Table 1, subjects who saw eight labeled cases learned enough from them to make 76.3% correct discriminations in Part II, many more than the no-labels group (58.5%). Their confidence was also higher, but with considerable justification. Again, subjects' learning ability thwarted our test of the hypothesis.

4. EXPERIMENT 3 - STOCKS

Experiment 3 replicated Experiments 1 and 2 with a task chosen to be truly impossible; predicting whether each of 12 common stocks had increased or decreased in price over the period from February 14, 1975, to March 19, 1975. The basis of these predictions was the stock market price and volume charts produced by Standard and Poor's Trendline division for the period July 12, 1974--February 14, 1975. Subjects first learned how to read the major features of such charts and then in Part I were allowed to study four charts of stocks, two of which had increased and two of which had decreased over the period. The labels group was told how these four stocks had performed in the next period; the no-labels group was not.

We had no reason to believe that this rudimentary training would enable people to predict market fluctuations (we would be in the wrong business if it did). Performance charts also appeared to be an attractive stimulus because many investors seem to stay in the market only because of their ability to create an illusion of explicability. Anyone who has heard even the brief stock market reports on the evening news knows that market analysts have an explanation for every fluctuation. Upon close examination, their explanatory processes seem to exemplify those described in our hypothesis. Analysts draw upon an enormous set of explanatory variables.[2] Not only is this set large enough to fit virtually any data, with a little ingenuity, but it contains contradictory explanatory rules. For example, if the market rises following good economic news, it is said to be responding to the news; if it falls, that is explained by saying that the good news had already been discounted. Figure 1 shows how two contradictory

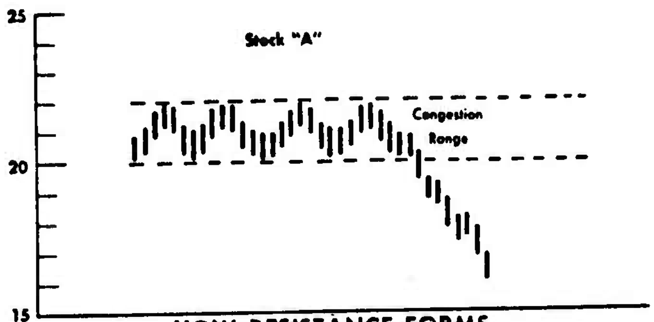
rules can be used, in hindsight, to show how a nondescript undulation in price foretold a subsequent increase or decrease in price (continued undulation, presumably, could be accounted for by a third rule).[3]

Whereas Fama (1965) has forcefully argued that market fluctuations are best understood as reflecting a random walk process, analysts' propensity for over-explaining is such that they seem to deny any random component in stock prices. Perhaps the best evidence of this is their reliance on the ultimate fudge factor for explaining random variations, the "technical adjustment."

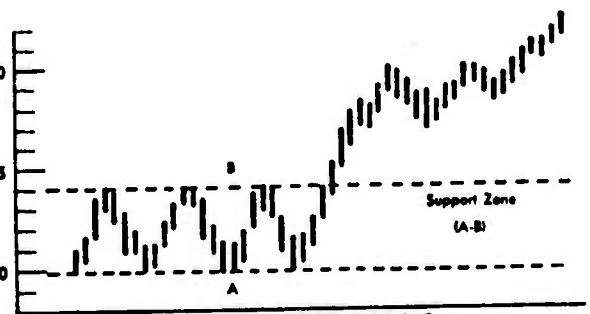
Method

Design. Experiment 3 replicated Experiments 1 and 2 except for the change of stimuli.

Stimuli. Four alternative sets of stimuli were created in the following manner: all 618 stocks appearing in Trendline for February 14, 1975, were sorted into those which were at least one point ($1) higher on March 19, 1975, those at least a point lower on March 19, and those which were relatively unchanged. For each of the four sets, two stocks showing increases were chosen to serve as study stimuli (Part I) and six more were chosen as test stimuli (Part II); two and six stocks showing decreases were also chosen. Stocks were chosen randomly without replacement. The relatively unchanged stocks were not used. Overall market indices were very similar on February 14 and March 19, indicating that there was no general market trend that knowledgeable subjects might use to improve their performance. A typical chart appears in Figure 2.



HOW RESISTANCE FORMS



HOW SUPPORT FORMS

Figure 1

Ambiguity in diagnostic signs (From W. Jiler, How Charts Can Help You in the Stock Market, New York: Trendline, 1962, used by permission).

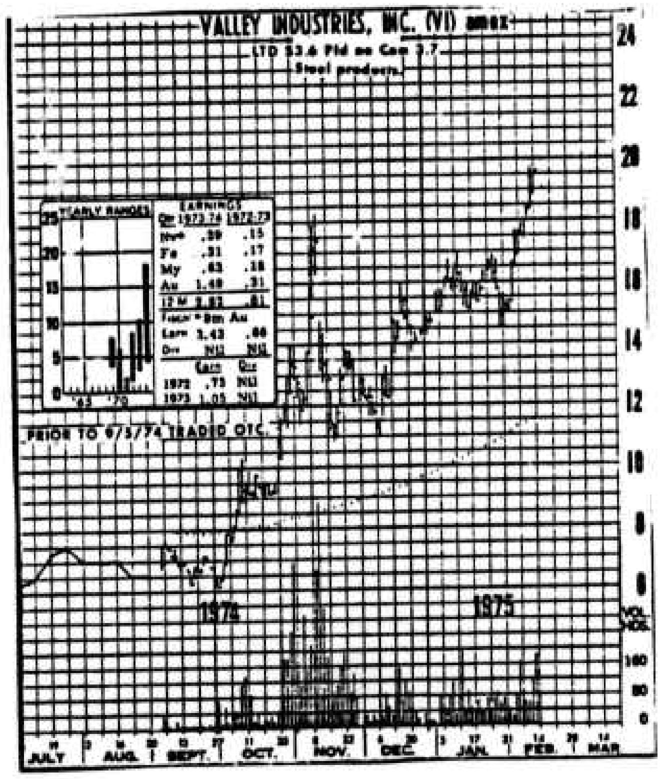


Figure 2

Typical stimulus for Experiment 3.

Procedure. A one-half hour explanation of how to read the Trendline charts was presented to the subjects. Questions were encouraged and answered to the group as a whole before proceeding to Parts I and II, which were analogous to the comparable sections of Experiment 1. A post-experiment questionnaire was used to identify subjects who had either specific knowledge of the stocks used or who had been totally confused by the task (there appeared to be none of either type) and to ask subjects about the strategies they had used.

Results

As hoped, labels subjects were unable to learn how to make the required discrimination. On Part II, they got only 49.3% correct. Nonetheless, they were substantially overconfident (mean probability = .643, overconfidence = .150).

Unfortunately for the hypothesis, no-labels subjects were just as confident (mean probability = .671) and, if anything, even more overconfident (percent correct = 44.1%, overconfidence = .230), without the benefit of labeled charts in Part I.

Discussion

The most dramatic result of Experiment 3 was the gross overconfidence of the no-labels subjects. Apparently, with only a brief explanation of how to read charts, these people believed that they were able to predict the direction of price movement for a variety of stocks. Given this initial overconfidence (which also characterized the no-labels group of Experiment 1), our manipulation would have had to be extremely powerful to have had any appreciable effect.

In exploring reasons for the no-labels subjects' overconfidence, we realized that the charts we were using also contained many labels for that group. They could, for example, generate labeled study trials by attempting to predict the February 14 closing price from that of January 9, or the February 13 close from that of January 8, and so on. In the post-experiment questionnaires, subjects in both groups reported basing their predictions on fairly elaborate rules, some drawn from their own intuitive theories of finance, others derived from study of the charts themselves. Given the amount of training information in the charts themselves, providing four March 19 closing prices to the labels group may have constituted a very minor addition.

5. EXPERIMENT 4 - HORSE RACING

The stock market task failed to test the "illusion of discriminability" hypothesis for two reasons: (1) no-labels subjects' undue confidence in their ability to perform an impossible task; and (2) the labels implicit in the stimuli given to no-labels subjects. Experiment 4 replicates the previous experiments with a task chosen to avoid these two problems: picking the winner from the first three horses in parimutuel races. We believed that no-labels subjects would see this as a task with a very large luck and a very small skill component, whereas possession of labels would lead subjects in the other group to the hypothesized overconfidence.

Method

Stimuli. Forty races held at the Aqueduct, New York, race track early in the 1968 season were selected from The Racing Form. The first three horses to finish from each race and 26 pieces of information about each horse were presented on a page like that in Table 2. Two paired sets of 10 races each were created out of the forty races. Each paired set was presented to half of the subjects, half of whom studied one member of the pair in Part I; the remaining subjects were tested on it in Part II. For the labels group, "winner" was written above the winning horse.

Instructions. All unfamiliar cues on the performance charts were explained to subjects in a group setting like that in Experiment 3. Instructions for Parts I and II were analogous to those in the previous experiments. In Part II, subjects were asked to choose the winning horse of the 3 and to give a probability ranging from 1/3 to 1.0 that they were correct.

Table 2

Typical Stimulus for Experiment 4

Name of Horse	Tillie's Alibi	Frostyann	Pookins
Age	5	4	4
Post Position	5	13	6
Modal Distance Raced	6f	6f	6f
1968: Number of Starts	4	6	4
1968: Number of Wins	0	0	0
1968: % Won	0	0	0
1968: Dollars Earned	500	2600	800
1967: Number of Starts	8	24	2
1967: Number of Wins	2	5	0
1967: % Won	25	21	0
1967 Dollars Earned	5800	22300	0
No. Days Since Last Race	10	10	47
Was Last Race at Aqueduct?	yes	yes	no
Finishing Position: Last Race	4	7	3
No. Lengths Behind in Last Race	-3.50	-8.0	-9.50
Speed Rating: Last Race	77	74	72
Weight This Race	116	116	116
Weight Last Race	114	115	113
Leading Jockey This Race?	yes	yes	yes
Jockey's 1967 Record: No. Starts	541	1648	388
Jockey's 1967 Record: No. Wins	32	301	28
Jockey's 1967 Record: % Won	6	18	7
Trainer's 1967 Record: No. Starts	76	393	263
Trainer's 1967 Record: No. Wins	5	39	31
Trainer's 1967 Record: % Won	6	10	12
Comment Last Race	Weakened	Bold bid,tired	Wide, tired

Results and Discussion

As Table 1 shows, both groups performed only slightly better than chance (33.3% correct) indicating the difficulty of the task both with and without labeled study examples. The marginal ability shown by all subjects was apparently due to several races where the winning horse clearly dominated the other two on the form charts. However, as in the previous experiments, subjects in <u>both</u> groups were grossly overconfident. Even without the benefit of labels, subjects believed that they could pick the winners. Again, the power of the experimental manipulation paled beside the strength of subjects' overconfidence.

6. EXPERIMENT 5 - CHILDREN'S DRAWINGS

Experiment 5 attempted to provide a fair test of the hypothesis by using a task that would appear patently impossible to no-labels subjects, allowing us (finally) to determine whether the presence of labels does make the impossible seem possible. The task chosen was sorting small children's drawings according to their continent of origin, Europe or Asia. These drawings were chosen from a book by Kellogg (1973), whose thesis is that the themes and forms of children's drawings are the same throughout the world.

Method

Stimuli. Each subject received in Part I a collection of 50-60 small (1-2 cm tall) children's drawings chosen from two European and two Asian countries which were either unlabeled (see Figure 3a) or labeled (see Figure 3b) according to their country of origin. All were taken from the inside front and back covers of Kellogg (1973). In Part II, subjects received, one at a time, 12 additional individual drawings (6 from each continent) that had been randomly selected from the drawings not used in Part I. Four sets of study and test drawings were compiled.

Instructions. Part I instructions for both groups were:

> In the present experiment, we are trying to determine whether people can discriminate between children's drawings from different parts of the world. In the first part of the experiment, you will have five minutes to familiarize yourself with sixty or so drawings of the type to be used on the second part. In that

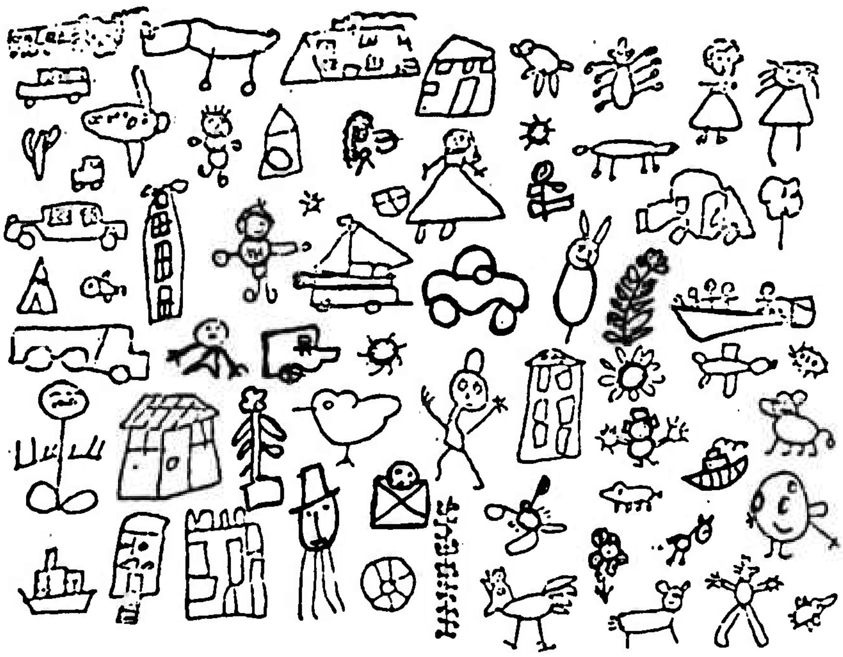


Figure 3a

Europe   Asia

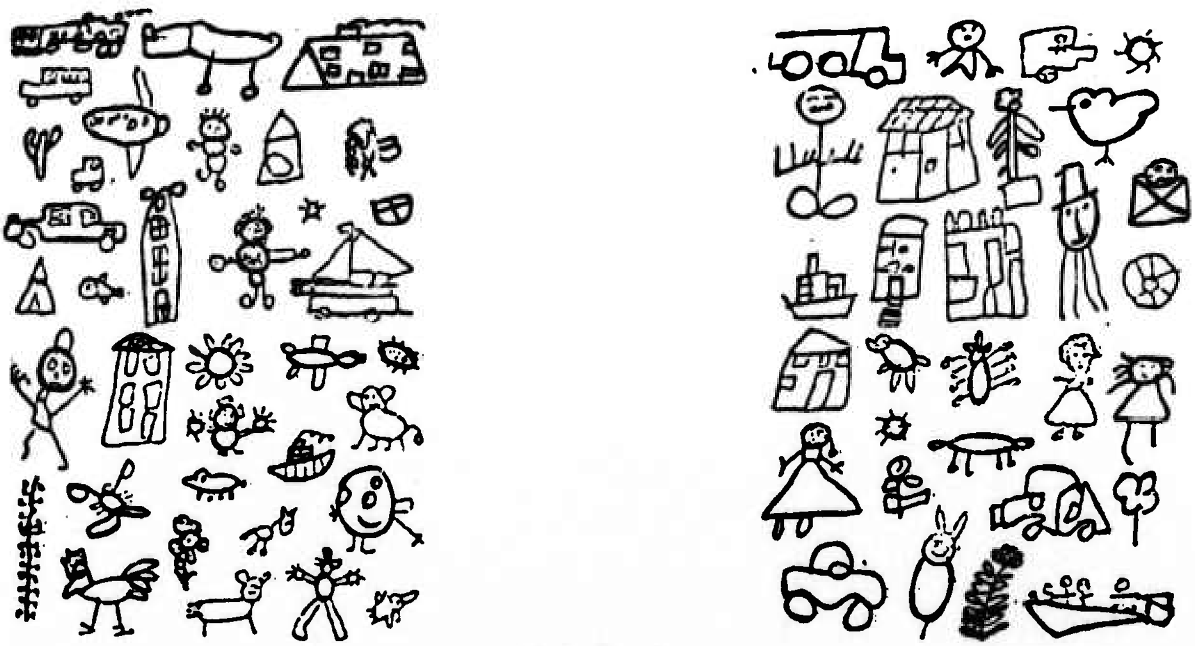


Figure 3b

Unlabeled (a) and labeled (b) study examples
for Experiments 5 and 6.

second part, you will be asked to decide for each of twelve drawings whether it comes from Europe or from South and East Asia. The European pictures all came from the following countries: Denmark, England, Germany, Greece, Italy, Spain, Sweden, or Switzerland. The South and East Asian drawings came from: China, Formosa, Hong Kong, India, Japan, Nepal, Philippines, or Thailand. All drawings were taken from the Rhoda Kellogg Child Art Collection.

Part II instructions were analogous to those used in previous tasks.

Results and Discussion

As Table 1 shows, the story of Experiments 3 and 4 was repeated. Labels subjects learned nothing by studying the labeled sketches. Both groups, however, were grossly overconfident. Apparently even this obscure task could not shake the no-labels subjects' confidence in their ability to make the required discriminations. Indeed, looking over the right-hand columns of Table 1, it appears that no-labels subjects give a mean probability response of about .65 regardless of the task and their ability to perform it.

Before concluding that this "65" rule is a cultural universal, it is worth considering the possibility that this overconfidence was induced, at least in part, by the instructions or experimental setting. In Experiments 1-5, care was taken to avoid any intimation that the task was possible so that the instructions would not be blamed for the anticipated overconfidence of labels subjects. Nonetheless, perhaps people

believe that any task set before them in an experiment must be possible. Experiment 6 was designed to reduce this possibility through the use of instructions stating explicitly that the children's drawings task might be impossible.

7. EXPERIMENT 6 - DISCOURAGING INSTRUCTIONS

Method

Instructions. The first sentence of the instructions for Experiment 5 was replaced with "Many people have claimed that the art of small children is the same in all cultural settings; others disagree. In the present experiment, we are trying to determine whether people can indeed discriminate between children's drawings from different parts of the world." The last sentence was replaced with "All drawings were taken from the Child Art Collection of Dr. Rhoda Kellogg, a leading proponent of the theory that children from different countries and cultures make very similar drawings." To the Part II instructions was appended "Remember, it may well be impossible to make this sort of discrimination. Try to do the best you can. But if, in the extreme, you feel totally uncertain about the origin of all of these drawings, do not hesitate to respond with .5 for every one of them."

Results

As Table 1 shows, the change in instructions had some effect in the appropriate direction, reducing the mean confidence of both groups by approximately .05. Both, however, were still overconfident. Only 6 of 76 subjects (4 in the labels group, 2 in the no-labels group) accepted the option of responding with .5 to all items (about the same proportion as in the previous experiments.)

8. EXPERIMENT 7 - BELLWETHER PRECINCTS

So far, we've learned more about the dangers of no learning than about the dangers of a little learning. Before abandoning our hypothesis, let us review the tasks we used to see whether, for all their variety, they might have shared some feature that kept labels subjects from capitalizing on chance correlations between independent variables and the dependent variable. One such common feature is the fact that the stimuli in all tasks were arranged by cases rather than by variables. To see if, for example, "number of days since last race" was a valid predictor of a winning horse, a subject would have to flip through 10 pages of races keeping a running tally of the correlation between that predictor (number of days) and the criterion. Keeping track of 26 such correlations and their relative strengths may have confused labels subjects and reduced their confidence. What would happen if our stimuli were organized by variables rather than cases or equally organized by both criteria? Except with horse racing, there is no way that any of the tasks we have used already could be so reorganized, in part because the potential predictors are not uniquely defined. One could not exhaustively list the characteristics of the children's drawings of Experiments 4 and 5. With horse racing, one could present each of the 26 predictors separately along with the horses and results from each of the 10 races. This arrangement would, however, eliminate the cases (races) as entities and present a highly unnatural array.

Experiment 7 explored the effect of organizing by predictors. Rather than rearrange the horse racing stimuli, we devised a new task allowing the stimuli to be organized either by cases or by predictors. In it, subjects were presented with

fictitious voting records for a number of precincts (4 or 8) over a number of elections (8 or 20) for one office. For each election and precinct, subjects were told which of the two parties running (D or R) was favored and by how much. Their task after studying the records was to predict the winning party on the next election on the basis of a pre-election poll of the precincts. The additional information given to labels subjects was who won each of the 8 or 20 study elections. In this task, the precincts are potential predictors and the election results are the criterion. Both the past election and pre-election poll results were generated randomly, so that there would, in fact, be no useful information for subjects to discern.

Method

Stimuli. Party preferences were generated using random normal deviates with a mean of 50 and a standard deviation of 12. The resulting numbers were treated as the percentage of voters favoring party D in each election. Numbers greater than 90 were treated as 90, those less than 10 were treated as 10. The results were presented in the form "party of preference, margin of victory." For example, a randomly generated number of 68 was interpreted as a vote of 68%d-32%R; it was presented to subjects as D-36 (=68-32). The election results were also generated randomly, with equal likelihood for both parties. All the election results were presented on one page of computer printout in one large matrix (see Figure 4). Election results (for labels subjects) appeared in separate lines above and below the matrix. Different subjects received different, independently generated matrices. Labels and no-labels subjects were yoked, each receiving the same matrix with pre-election poll results. However, only labels subjects saw the election

SUBJECT #2

	ELECTION			
	1	2	3	4
Winner	D	D	R	R
Precinct #				
1	R 37	R 13	R 5	D 3
2	R 23	R 41	R 12	R 27
3	D 12	R 59	D 15	D 38
4	D 1	R 2	R 14	R 39
5	R 13	R 17	D 12	D 4
6	R 6	D 13	R 41	R 17
7	R 27	D 8	R 23	D 29
8	R 14	R 23	D 25	R 43

Figure 4

Typical study example for Experiment 7.

results. Three matrix sizes were used: (a) 8 elections and 20 precincts; (b) 4 elections and 20 precincts; and (c) 4 elections and 8 precincts.

Procedure. Subjects studied the matrix for 10 minutes after being told:

> Are there bellwether electoral precincts, precincts on the basis of whose voting record we can predict the outcome of future elections? Some people believe there are; others disagree. In the present experiment, we are trying to determine whether people can predict the outcome of a future election on the basis of the voting record of several randomly selected precincts.

After their study period, subjects were presented pre-election poll results for that next election and were asked to (1) predict the winner of that election and (2) indicate their confidence in having picked the winner. Confidence was elicited in odds rather than probabilities. In other experiments (Fischhoff, Slovic & Lichtenstein, 1977), we had found that odds judgments are less likely than probability judgments to be rounded to a few stereotypic responses (.50, .60, .70, etc.). We hoped that using odds would provide greater sensitivity.

Results and Discussion

As Table 3 shows, there was not consistent pattern of results. For the [8 elections, 20 precincts] condition, the labels group gave greater median odds that their predictions were correct; for the 4 x 8 condition, they gave smaller median odds; for the 4 x 20 condition, the median odds for the groups were about equal. None of these differences were statistically significant (median test; alpha = .05). Analyses

Table 3

Bellwether Precincts -- Experiment 7

Median Odds of Being Correct

	# of Elections	4	4	8
	# of Precincts	8	20	20
No Labels Group		5	2.5	2
		(56)	(28)	(35)
Labels Group		2	2	3
		(59)	(33)	(38)

Note: Number of subjects appears in parentheses.

done in terms of yoked labels and no-labels subjects (who saw the same randomly generated matrix and election poll results) also showed no consistent differences.[4]

What went wrong this time? The most parsimonious explanation in light of the earlier results is that as soon as they were confronted with the task, no-labels subjects felt an undue confidence in their own abilities. The labels manipulation was an inconsequential factor compared to this overconfidence. Two additional factors may have weakened the design of this particular experiment. One is that some no-labels subjects created their own labels by totaling the results in the precincts presented on the study elections and treating those as total election results. Explicit totaling could be seen on the forms of about a quarter of the no-labels subjects. The second problem is that a portion of subjects apparently found the task of pouring through a large matrix of numbers quite frustrating and "gave up."

9. EXPERIMENT 8 - AMOUNT OF STUDY

Two aspects of these results need explaining: (1) why are no-labels subjects so confident? and (2) why doesn't the addition of labels induce even more confidence?

People's overconfidence in their general knowledge and intellectual ability is apparently a widespread and robust tendency (Fischhoff, Slovic & Lichtenstein, 1977; Slovic, Fischhoff & Lichtenstein, 1977, pp. 5-6, 14-17). When called upon to answer a particular question, people seem unaware of the tenuousness of their reasoning and assumptions or of the contrary evidence they have overlooked. When confronted with a series of similar tasks, many people may also generate an inappropriate global feeling of confidence: "Here's a task I can handle." This feeling may come from personal experience with a related task ("I've done quite a bit of handwriting analysis in the past") or from a culturally shared belief that the task (any task?) is tractable given the proper information (e.g., "One can win at the races with proper research" or "There are bellwether precincts to be found if one looks hard enough"--however, see Tufte & Sun, 1975, for evidence to the contrary). Although we tried not to encourage such expectations (especially in Experiment 6), nothing short of telling subjects that the task is impossible may be adequate.

One reason why the addition of labeled feedback may not augment this overconfidence is the fairly large number of study trials with which subjects were confronted. Finding one cue or a combination of cues that discriminate the two sets of stimuli for each of 10 to 12 trials may not be easy. Depending on how quickly they complete the search, subjects might realize the element of luck in their success or, more likely,

just feel that the task was harder than it looked. For example, they may discover that cues that a priori they would have expected to discriminate do not. The reduction in confidence arising from discovering such difficulties may cancel the increase in confidence arising from discovery of a rule. Reducing the number of study trials will increase the likelihood that some cues will be perfectly consistent discriminators and, therefore, may lead to labels groups that are more confident. Experiment 8 explores this possiblity by presenting a minimal number of study examples to labels subjects.

In both Experiment 3 (Stocks) and Experiment 7 (Bellwether precincts), we found evidence that some subjects in the no-labels group were, quite ingeniously, producing their own labels. We suspect that some form of self-generated feedback may be quite common. For example, no-labels subjects might decide that some handwriting samples look American while others look European, and then set out to figure out why. In doing so, they may not only be converting their task to that of labels subjects, but doing so in a way that makes finding a good discriminatory rule quite easy: for one, they may be considering a reduced set of trial samples (those that appear clear-cut examples of one category or the other). In addition, their validation process may be circular. They may start out with one or several cues that seem a priori to be valid, use them to pick clear-cut cases, and then validate the cues by how well they work on the selected cases. In such a situation, a cue seems valid if it can be applied. Eliminating such self-generated cue validation would seem to be quite difficult. Experiment 8 tries to do so by eliminating the study session entirely. No-labels subjects went directly to the test examples of Part II.

Method

Design. Two new versions were created for four of the tasks used in previous experiments. One version contained a minimal number of labeled study examples; the other contained no study section at all. The test examples of these tasks were identical to those used earlier. The tasks used were handwriting (Experiment 1), ulcers (Experiment 2), horse racing (Experiment 4) and children's drawings (discouraging instructions version--Experiment 6). Stocks and bellwether precincts were not used again because they were found to contain implicit feedback which was noted and exploited by some subjects. Handwriting and ulcers were used with some trepidation since the labels subjects in Experiments 1 and 2 were able to improve their performance on the basis of what they learned in the study section. It was hoped that the abbreviated study session given the present labels group would not provide such an opportunity for learning.

Stimuli. For the handwriting task, abbreviated versions of the study session (Part I) were created by using one European and one American handwriting sample (both labeled). For ulcers, the abbreviated study session contained one benign and one malignant example (labeled). For horse racing, two races were presented with the winners indicated. Children's drawings subjects saw five European and five Asian examples used in the abbreviated study session which were drawn randomly from those used in the full study sessions. Several such samples were drawn from each Part I and used with a portion of the subjects. For the no-study condition, tasks were created by combining those sections of the Part I instructions explaining the task with Part II instructions.

Subjects. Three hundred and thirty-three subjects were recruited as before.

Results

No study session. As the right half of Table 4 shows, eliminating the study session entirely had no systematic effect on no-labels subjects. With handwriting, horse racing and children's drawings, mean confidence and percent correct were virtually the same for the present subjects and those shown 10 unlabeled examples. With ulcers, percent correct went down somewhat and confidence increased, suggesting that the minimal overconfidence (.014) observed in Experiment 2 was only a fluke.

Abbreviated labeled study sessions. Remarkably, seeing one pair of labeled examples enabled both handwriting and ulcers subjects to perform somewhat better than chance. They were more confident than the corresponding no-labels subjects (who did no better than chance), but this increased confidence was justified. The horse racing and children's drawings groups provide a better test of the effect of worthless study on confidence, since the few labeled examples they saw did not improve their performance. Their mean confidence was indistinguishable from that of subjects who studied 10 labeled examples.

Table 4

Experiment 8: Amount of Study

Number of Cases Studied	Labels: Percent Correct	Labels: Mean Probab.	Labels: Over-Under Confidence[a]	Labels: N	No Labels: Percent Correct	No Labels: Mean Probab.	No Labels: Over-Under Confidence	No Labels: N
Handwriting								
10(Exp. 1)	77.0	.745	-.025	22	53.3	.645	.112	30
2	62.9	.705	.076	45				
0					56.8	.641	.073	40
Ulcers								
8(Exp. 2)	76.3	.702	-.061	33	58.5	.599	.014	38
2	70.5	.673	-.033	42				
0					50.0	.643	.143	39
Horse Racing								
10(Exp. 4)	41.5	.603	.188	46	39.1	.651	.260	42
2	40.7	.624	.217	44				
0					40.0	.621	.221	38
Children's Drawings (Discouraging Instructions)								
60(Exp. 6)	57.7	.631	.054	40	45.6	.627	.171	36
10	51.9	.651	.132	44				
0					51.1	.650	.139	41

[a] Equals difference between mean probability and percent correct. Negative sign indicates underconfidence.

10. CONCLUDING DISCUSSION

Using a variety of tasks, instructions and study sessions, these experiments have confirmed the most robust result of previous work on confidence (Fischhoff, Slovic & Lichtenstein, 1977; Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff & Phillips, 1977): people are consistently overconfident in their ability to perform difficult or impossible tasks with which they have some minimal familiarity. As performance improves, overconfidence decreases.

Our attempts to manipulate confidence through the provision of useless study examples were humbled by this imported overconfidence. The fact that subjects were just as confident in the absence of study sessions (Experiment 8) as with them suggests that mere exposure to a comprehensible task leads people to feel that they have some competence to perform it. Some possible reasons for this illusion of competence were discussed earlier. Perhaps the most interesting explanation to receive support from these studies is that confidence may be relatively independent of immediate experience. It would seem as though the very ability to generate an applicable rule from discrimination carries with it a conviction that the rule has some validity. Since it is almost always possible to generate some rule (e.g., "'rugal pattern' sounds malignant to me") overconfidence should then be the rule rather than the exception.

Once generated, confidence may be very difficul' to dispel, for it is unusual to receive a concentrated set of clearly labeled examples of the sort needed to test one's rules (Goldberg, 1967; Skinner, 1968). More typically, such feedback as we receive is late (so that we forget or misremember

our predictions), spread over time (so that its cumulative impact is lost), or ambiguous (so that we can explain away our mistakes). All these characteristics of our experience could tend to leave our confidence unshaken by experience. And, on those rare occasions when feedback is prompt and precise, we may not know how to use it to assess discriminability (Wason & Johnson-Laird, 1972; Einhorn & Hogarth, in press).

How has the present concentrated, immediate and unambiguous experience affected our confidence in the hypothesis that motivated this enterprise? Rather little. We still believe that capitalization on chance patterns can generate undue confidence in erroneous theories. What has changed is our belief in the prevalence of looking for patterns as a mode of learning and determining confidence. Although an effective path to overconfidence, capitalization upon chance may not be a necessary one.

11. REFERENCES

Armstrong, J.S. Tom Swift and his electric regression analysis machine: 1973. Psychological Reports, 1975, 36: 806.

Campbell, D.T. Degrees of freedom and the case study. Comparative Political Studies, 1975, 8: 2.

Crask, M.R. & Perreault, W.D., Jr. Validation of discriminant analysis in marketing research. Journal of Marketing Research, 1977, 14: 60-68.

Einhorn, J.H. & Hogarth, R.M. Confidence in judgment: Persistence of the illusion of validity. Psychological Review, in press.

Fama, E.F. Random walks in stock market prices. Financial Analysts Journal, 1965, 21: 55-60

Fischhoff, B. Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1975, 1: 288-299.

Fischhoff, B. Perceived informativeness of facts. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3: 349-358.

Fischhoff, B., Slovic, P. & Lichtenstein, S. Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3: 552-564.

Hammond, K.R. & Summers, D.A. Cognitive control. Psychological Review, 1972, 79: 58-67.

Hamner, W.C. The importance of sample size, cut-off technique, and cross-validation in multiple regression analysis. Decision Science, 1975.

Kellogg, R. Analyzing Children's Art. Palo Alto, Calif.: National Press, 1970.

Kunce, J.T., Cook, D.W. & Miller, D.E. Random variables and correlational overkill. Educational and Psychological Measurement, 1975, 35: 529-534.

Langer, E.J. The illusion of control. Journal of Personality and Social Psychology, 1975, 32: 311-328.

Langer, E.J. The psychology of chance. Journal for the Theory of Social Behavior, 1977, 7: 185-208.

Lewis-Beck, M.S. Stepwise regression: A caution to users. 1977 Annual Meeting of the Midwest Political Science Association, The Pick-Congress Hotel, Chicago, Ill., April 21-23, 1977.

Lichtenstein, S. & Fischhoff, B. Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance, 1977, 20: 159-183.

Lichtenstein, S., Fischhoff, B. & Phillips, L.D. Calibration of probabilities: The state of the art. In J. Jungermann & G. de Zeeuw (Eds.), Decision Making and Change in Human Affairs. Amsterdam: D. Reidel, 1977.

O'Leary, M.K., Coplin, W.D., Shapiro, H.B. & Dean, D. The quest for relevance. International Studies Quarterly, 1974, 18: 211-237.

Slovic, P., Fischhoff, B. & Lichtenstein, S. Behavioral decision theory. Annual Review of Psychology, 1977, 28: 1-39.

Slovic, P., Rorer, L. & Hoffman, P.J. Analyzing the use of diagnostic signs. Investigative Radiology, 1971, 6: 18-26.

Tufte, E.R. & Sun, R.A. Are there bellwether electoral districts? The Public Opinion Quarterly, 1975, 39: 1-18.

Wason, P.C. & Johnson-Laird, P.N. Psychology of Reasoning: Structure and Content. London: D.T. Batsford, 1972.

12. FOOTNOTES

1. O'Leary, Coplin, Shapiro and Dean (1974), in a study of the explanatory protocols used by U.S. Department of State foreign affairs analysts, found that analysts relied on multivariate, explanatory models using discrete variables with nonlinear, time-lagged relationships between them. They observed that "the kinds of relationships found in the majority of [State Department] analyses represent such complexity that no single quantitative work in the social sciences could even begin to test their validity" (p. 228).

2. One of the authors once took a course in reading form charts from a local brokerage. Each session involved the teaching of 10-12 new cues. When the course ended, 8 sessions and 83 cues later, the instructor was far from exhausting his supply.

3. Exploitation of the ambiguity of such signs to make contradictory forecasts may be seen in the following quote from Business Week. "[A well-known economist] translates these pressures into an inflation rate of 8% to 9% by the final quarter of this year. And those numbers are springs on a bear trap, unless Wall Street has once again decided that inflation is good for stock prices" (May 8, 1978, p. 28).

4. Not only are these results disappointing, the weak interaction exhibited in Table 3 actually goes somewhat in the opposite direction from what one might expect. Reducing the number of elections from 8 to 4 (while holding the number of precincts constant at 20) increases the probability of there being at least one bellwether precinct (predicting the results

of all elections correctly) from .07 to .33. In addition, reducing the number of elections made the whole task considerably easier, increasing labels subjects' chances of finding a bellwether precinct if one were present. Nonetheless, labels subjects were relatively less confident in the 4 x 20 condition than in the 8 x 20 condition.

5. A horse racing group that has two unlabeled examples was also conducted (N = 44). They showed about the same percentage correct (37.3%), mean confidence (.623) and overconfidence (.250) as the other horse racing groups.

CONTRACT DISTRIBUTION LIST
(Unclassified Technical Reports)

Director Advanced Research Projects Agency Attention: Program Management Office 1400 Wilson Boulevard Arlington, Virginia 22209	2 copies
Office of Naval Research Attention: Code 455 800 North Quincy Street Arlington, Virginia 22217	3 copies
Defense Documentation Center Attention: DDC-TC Cameron Station Alexandria, Virginia 22314	12 copies
DCASMA Baltimore Office Attention: Mr. K. Gerasim 300 East Joppa Road Towson, Maryland 21204	1 copy
Director Naval Research Laboratory Attention: Code 2627 Washington, D.C. 20375	6 copies
Decisions and Designs, Incorporated 8400 Westpark Drive, P.O. Box 907 McLean, Virginia 22101	10 copies

SUPPLEMENTAL DISTRIBUTION LIST
(Unclassified Technical Reports)

Department of Defense

Director of Net Assessment
Office of the Secretary of Defense
Attention: MAJ Robert G. Gough, USAF
The Pentagon, Room 3A930
Washington, DC 20301

Assistant Director (Net Technical Assessment)
Office of the Deputy Director of Defense
Research and Engineering (Test and
Evaluation)
The Pentagon, Room 3C125
Washington, DC 20301

Director, Defense Advanced Research
Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, ARPA Regional Office (Europe)
Headquarters, U.S. European Command
APO New York 09128

Director, ARPA Regional Office (Pacific)
Staff CINCPAC, Box 13
APO San Francisco 96610

Dr. Don Hirta
Defense Systems Management School
Building 202
Ft. Belvoir, VA 22060

Chairman, Department of Curriculum
Development
National War College
Ft. McNair, 4th and P Streets, SW
Washington, DC 20319

Defense Intelligence School
Attention: Professor Douglas E. Hunter
Washington, DC 20374

Vice Director for Production
Management Office (Special Actions)
Defense Intelligence Agency
Room 1E863, The Pentagon
Washington, DC 20301

Command and Control Technical Center
Defense Communications Agency
Attention: Mr. John D. Hwang
Washington, DC 20301

Department of the Navy

Office of the Chief of Naval Operations
(OP-951)
Washington, DC 20450

Office of Naval Research
Assistant Chief for Technology (Code 200)
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Director, ONR Branch Office
Attention: Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
Attention: Dr. M. Bertin
American Embassy - Room A-407
APO San Francisco 96503

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps (Code RD-1)
Washington, DC 20380

Headquarters, Naval Material Command
(Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Dean of Research Administration
Naval Postgraduate School
Attention: Patrick C. Parker
Monterey, CA 93940

Superintendent
Naval Postgraduate School
Attention: R. J. Roland, (Code 52Rl)
C^3 Curriculum
Monterey, CA 93940

Naval Personnel Research and Development
Center (Code 305)
Attention: LCDR O'Bar
San Diego, CA 92152

Navy Personnel Research and Development
Center
Manned Systems Design (Code 311)
Attention: Dr. Fred Muckler
San Diego, CA 92152

Naval Training Equipment Center
Human Factors Department (Code N215)
Orlando, FL 32813

Naval Training Equipment Center
Training Analysis and Evaluation Group
(Code N-00T)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Director, Center for Advanced Research
Naval War College
Attention: Professor C. Lewis
Newport, RI 02840

Naval Research Laboratory
Communications Sciences Division (Code 5403
Attention: Dr. John Shore
Washington, DC 20375

Dean of the Academic Departments
U.S. Naval Academy
Annapolis, MD 21402

Chief, Intelligence Division
Marine Corps Development Center
Quantico, VA 22134

Department of the Army

Deputy Under Secretary of the Army
(Operations Research)
The Pentagon, Room 2E621
Washington, DC 20310

Director, Army Library
Army Studies (ASDIRS)
The Pentagon, Room 1A534
Washington, DC 20310

U.S. Army Research Institute
Organizations and Systems Research Laboratory
Attention: Dr. Edgar M. Johnson
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and Systems
Research Laboratory
U.S. Army Institute for the Behavioral
and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Technical Director, U.S. Army Concepts
Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20014

Director, Strategic Studies Institute
U.S. Army Combat Developments Command
Carlisle Barracks, PA 17013

Commandant, Army Logistics Management Center
Attention: DRXMC-LS-SCAD (ORSA)
Ft. Lee, VA 23801

Department of Engineering
United States Military Academy
Attention: COL A. F. Grum
West Point, NY 10996

Marine Corps Representative
U.S. Army War College
Carlisle Barracks, PA 17013

Chief, Studies and Analysis Office
Headquarters, Army Training and Doctrine
Command
Ft. Monroe, VA 23351

Commander, U.S. Army Research Office
(Durham)
Box CM, Duke Station
Durham, NC 27706

<u>Department of the Air Force</u>

Assistant for Requirements Development
and Acquisition Programs
Office of the Deputy Chief of Staff for
Research and Development
The Pentagon, Room 4C331
Washington, DC 20330

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Commandant, Air University
Maxwell AFB, AL 36112

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Deputy Chief of Staff, Plans, and
Operations
Directorate of Concepts (AR/XOCCC)
Attention: Major R. Linhard
The Pentagon, Room 4D 1047
Washington, DC 20330

Director, Advanced Systems Division
(AFHRL/AS)
Attention: Dr. Gordon Eckstrand
Wright-Patterson AFB, OH 45433

Commander, Rome Air Development Center
Attention: Mr. John Atkinson
Griffis AFB
Rome, NY 13440

IRD, Rome Air Development Center
Attention: Mr. Frederic A. Dion
Griffis AFB
Rome, NY 13440

HQS Tactical Air Command
Attention: LTCOL David Dianich
Langley AFB, VA 23665

Other Government Agencies

Chief, Strategic Evaluation Center
Central Intelligence Agency
Headquarters, Room 2G24
Washington, DC 20505

Director, Center for the Study of Intelligence
Central Intelligence Agency
Attention: Mr. Dean Moor
Washington, DC 20505

Mr. Richard Heuer
Methods & Forecasting Division
Office of Regional and Political Analysis
Central Intelligence Agency
Washington, DC 20505

Office of Life Sciences
Headquarters, National Aeronautics and Space Administration
Attention: Dr. Stanley Deutsch
600 Independence Avenue
Washington, DC 20546

Other Institutions

Department of Psychology
The Johns Hopkins University
Attention: Dr. Alphonse Chapanis
Charles and 34th Streets
Baltimore, MD 21218

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

Director, Social Science Research Institute
University of Southern California
Attention: Dr. Ward Edwards
Los Angeles, CA 90007

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364 (10 copies)

Stanford University
Attention: Dr. R. A. Howard
Stanford, CA 94305

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A. D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Brunel University
Attention: Dr. Lawrence D. Phillips
Uxbridge, Middlesex UB8 3PH
England

Decision Analysis Group
Stanford Research Institute
Attention: Dr. Miley W. Merkhofer
Menlo Park, CA 94025

Decision Research
1201 Oak Street
Eugene, OR 97401 (10 copies)

Department of Psychology
University of Washington
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Department of Electrical and Computer Engineering
University of Michigan
Attention: Professor Kan Chen
Ann Arbor, MI 94135

Department of Government and Politics
University of Maryland
Attention: Dr. Davis B. Bobrow
College Park, MD 20747

Department of Psychology
Hebrew University
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. Andrew P. Sage
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22901

Professor Raymond Tanter
Political Science Department
The University of Michigan
Ann Arbor, MI 48109

Professor Howard Raiffa
Morgan 302
Harvard Business School
Harvard University
Cambridge, MA 02163

Department of Psychology
University of Oklahoma
Attention: Dr. Charles Gettys
455 West Lindsey
Dale Hall Tower
Norman, OK 73069

Institute of Behavioral Science #3
University of Colorado
Attention: Dr. Kenneth Hammond
Room 201
Boulder, Colorado 80309

unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER PTR-1060-78-6	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Little Learning . . . : Confidence in Multicue Judgment Tasks		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Baruch Fischhoff and Paul Slovic		8. CONTRACT OR GRANT NUMBER(s) ARPA N00014-78-C-0100
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Research A Branch of Perceptronics 1201 Oak Street, Eugene, Oregon 97401		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order No. 3469
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, Virginia 22217		12. REPORT DATE June 1978
		13. NUMBER OF PAGES 56
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217		15. SECURITY CLASS. (of this report) unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

none

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

confidence
calibration
multicue discrimination
decision making
uncertainty

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A variety of discrimination tasks using complex, multifaceted stimuli were presented to subjects either with or without the opportunity to study a number of labeled examples. These tasks included deciding whether handwriting samples were produced by an American or a European, whether an ulcer was benign or malignant and which of three horses was a winner of a race at Aqueduct in 1969. Complex stimuli were chosen so that there would be a high probability that in the labeled study examples, diligent subjects could find some cue(s) highly

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

6 9 10 15 11 12 58p.

390 664

1B

correlated with the labels. Such capitalization on chance correlations has often been cited as the source of scientists' unwarranted confidence in their theories. As anticipated, subjects who studied examples were consistently overconfident. However, subjects who studied unlabeled examples or no examples at all were equally overconfident. Some reasons for the independence of confidence from immediate experience are discussed.